Quick Find an Author:

Enter a last name to quickly locate articles by that author.

Go

Note: You may enter a partial name if your are unsure of the spelling.
OR
Select a letter to browse the author list

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z | ALL

Tresch M. Trescher J.

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z | ALL

Your search matched **3** of **981130** documents.
Results are shown **15** to a page, sorted by **publication year** in **descending** order.

Results:
Journal or Magazine = **JNL** Conference = **CNF** Standard = **STD**

1 **Towards heterogeneous multimedia information systems: the Garlic approach**
Carey, M.J.; Haas, L.M.; Schwarz, P.M.; Arya, M.; Cody, W.F.; Fagin, R.; Flickner, M.; Luniewski, A.W.; Niblack, W.; Petkovic, D.; Thomas, J.; Williams, J.H.; Wimmers, E.L.;
Research Issues in Data Engineering, 1995: Distributed Object Management, Proceedings. RIDE-DOM '95. Fifth International Workshop on , 6-7 March 1995
Page(s): 124 -131

[Abstract] [PDF Full-Text (772 KB)] **IEEE CNF**

2 **A design for fine-grained access control in Melampus**
Luniewski, A.W.; Stamos, J.W.; Cabrera, L.-F.;
Object Orientation in Operating Systems, 1991. Proceedings., 1991 International Workshop on , 17-18 Oct. 1991
Page(s): 185 -189

[Abstract] [PDF Full-Text (328 KB)] **IEEE CNF**

3 **Quill: an extensible system for editing documents of mixed type**
Chamberlin, D.D.; Hasselmeier, H.F.; Luniewski, A.W.; Paris, D.P.; Wade, B.W.; Zolliker, M.L.;
System Sciences, 1988. Vol.II. Software Track, Proceedings of the Twenty-First Annual Hawaii International Conference on , Volume: 2 , 5-8 Jan. 1988
Page(s): 317 -326

[Abstract] [PDF Full-Text (784 KB)] **IEEE CNF**

Your search matched **2** of **981130** documents.
Results are shown **15** to a page, sorted by **publication year** in **descending** order.

Results:
Journal or Magazine = **JNL** Conference = **CNF** Standard = **STD**

1 **A classification of multi-database languages**
Tresch, M.; Scholl, M.H.;
Parallel and Distributed Information Systems, 1994., Proceedings of the Third International Conference on , 28-30 Sept. 1994
Page(s): 195 -202

[Abstract] [PDF Full-Text (668 KB)] **IEEE CNF**

2 **The FEMUS approach in building a federated multilingual database system**
Andersson, M.; Dupont, Y.; Spaccapietra, S.; Yetongnon, K.; Tresch, M.; Ye, H.;
Research Issues in Data Engineering, 1993: Interoperability in Multidatabase Systems, 1993. Proceedings RIDE-IMS '93., Third International Workshop on , 19-20 April 1993
Page(s): 65 -68

[Abstract] [PDF Full-Text (344 KB)] **IEEE CNF**

Google Advanced Search Preferences Language Tools Search Tips

+"type classification of semi-structured documents" Google Search

Web Images Groups Directory News

Searched the web for +"type classification of semi-structured documents". Results 1 - 10 of about 67. Search

[PDF] **Type Classification of Semi-Structured Documents**
File Format: PDF/Adobe Acrobat - View as HTML
Page 1. Page 2. Page 3. Page 4. Page 5. Page 6. Page 7. Page 8. Page 9. Page 10. Page 11. Page 12.
www.acm.org/sigmod/vldb/conf/1995/P263.PDF - Similar pages

VLDB 1995: 263-274
Type Classification of Semi-Structured Documents. Markus Tresch, Neal Palmer, Allen Luniewski: **Type Classification of Semi-Structured Documents**. ...
www.informatik.uni-trier.de/~ley/ db/conf/vldb/TreschPL95.html - 14k - Nov 1, 2003 - Cached - Similar pages

CIKM 1995: 226-233
... CACM 18(11): 613-620(1975) [TPL94] Markus Tresch, Neal Palmer, Allen Luniewski: **Type Classification of Semi-Structured Documents**. ...
www.informatik.uni-trier.de/~ley/ db/conf/cikm/TreschL95.html - 12k - Cached - Similar pages
[More results from www.informatik.uni-trier.de]

An extensible classifier for semi-structured documents
... 1975. TPL94 M. Tresch, N. Palmer, and A. Luniewski. **Type classification of semi-structured documents**, in Proc. 21th Int'l Conf. on ...
portal.acm.org/ citation.cfm?id=221575&jmp=references&dl=portal&dl=ACM&CFID=11111111&C... - Similar pages

Generating Association Rules from Semi-Structured Documents Using ...
... (context) - Klemettined, Mannila et al. - 1994 5 **Type Classification of Semi-structured Documents** (context) - Trensh, Palmer et al. ...
citeseer.nj.nec.com/singh97generating.html - 25k - Cached - Similar pages

A Robust System Architecture for Mining Semi-structured Data - ...
... Documents .. - Singh, Scheuermann et al. - 1997 5 **Type Classification of Semi-structured Documents** (context) - Tresch, Palmer et al. - 1995 ...
citeseer.nj.nec.com/singh98robust.html - 24k - Cached - Similar pages
[More results from citeseer.nj.nec.com]

[PDF] A Robust System Architecture for Mining Semi-structured Data
File Format: PDF/Adobe Acrobat - View as HTML
... Las Vegas, Nev.: ACM Press. Tresch, M., Palmer, N., and Luniewski, A. 1995. **Type Classification of Semi-structured Documents**. In ...
www.ece.northwestern.edu/EXTERNAL/ dbwww/papers/KDD98.pdf - Similar pages

Seminar SS99 - [Translate this page]
... SIGMOD 1998, 307-318. [5] Markus Tresch, Neal Palmer, Allen Luniewski: **Type Classification of Semi-Structured Documents**. VLDB 1995, 263-274. ...
www.informatik.uni-stuttgart.de/ipvr/ as/lehre/seminar/seminar_ss99.html - 11k - Cached - Similar pages

DBLP: Markus Tresch
... CIKM 1995: 226-233. 13, EE, Markus Tresch, Neal Palmer, Allen Luniewski:

Type Classification of Semi-Structured Documents. VLDB 1995: 263-274. 1994. ...
dblab.comeng.cnu.ac.kr/~dolphin/db/indices/ a-tree/t/Tresch:Markus.html - 9k - Cached - Similar pages

[PS] Ontology-Based Binary-Categorization of Multiple-Record Web ...
File Format: Adobe PostScript - View as Text
... 15. [TPL95] Markus Tresch, neal Palmer, Allen Luniewski. **Type Classification of Semi-Structured Documents**. Proceeding of the 21th VLDB Conference. ...
www.deg.byu.edu/proposals/ThesisPro.wang.ps - Similar pages

Gooogle ▶
Result Page: 1 2 3 **Next**

Google Advanced Search Preferences Language Tools Search Tips

+"type classification of semi-structured documents" Google Search

Web Images Groups Directory News

Searched the web for **+"type classification of semi-structured documents"**. Results **11** - **18** of about **67**. Searc

[PDF] Recognizing Ontology-Applicable Multiple-Record Web Documents
File Format: PDF/Adobe Acrobat - View as HTML
Page 1. Recognizing Ontology-Applicable Multiple-Record Web Documents
DW Embley, YK Ng, L. Xu Department of Computer Science Brigham ...
www.deg.byu.edu/papers/er01.pdf - Similar pages

@PROCEEDINGS{VLDB95, TITLE = {Proceedings of the 21st ...
... InProceedings{VLDB95*263, author = {Tresch, M. and Palmer, N. and Luniewski, A.},
title = {**Type Classification of Semi-Structured Documents**}, pages = {263--274 ...
www.mpi-sb.mpg.de/~elidee/toc/ proceedings/fb14/vldb95.bib - 15k - Cached - Similar pages

<html> <head> </head><body><pre><html> <head> </ ...
... Session 8: Multimedia WE 13 09, 14:00 - 15:30 - M. Tresch, N. Palmer, A. Luniewski
(USA): **Type Classification of Semi-Structured Documents** - F. Moser, A. Kraiss ...
www.lirmm.fr/~reitz/conferences/1995/VLDB.txt - 56k - Cached - Similar pages

VLDB 1993: 97-107
... Engine. ICDE 1996: 172-179; Markus Tresch, Neal Palmer, Allen Luniewski:
Type Classification of Semi-Structured Documents. VLDB 1995 ...
ftp.informatik.uni-trier.de/~ley/ db/conf/vldb/SoensLSST93.html - 22k - Supplemental Result - Cached - Similar pages

ICDT 1997: 1-18
... VLDB 1996: 227-238 [TMD92] ... [TPL95] Markus Tresch, Neal Palmer, Allen
Luniewski: **Type Classification of Semi-Structured Documents**. ...
www.sigmod.org/sigmod/dblp/db/ conf/icdt/Abiteboul97.html - 41k - Supplemental Result - Cached - Similar pages

SIGMOD Conference 1994: 313-324
... VLDB Journal 4(1): 45-86(1995); Markus Tresch, Neal Palmer, Allen Luniewski:
Type Classification of Semi-Structured Documents. VLDB ...
dbweb.csie.ncu.edu.tw/DBLP/dblp/db/conf/ sigmod/ChristophidesACS94.html - 31k - Supplemental Result - Cached - Similar pages

SIGMOD Conference 1994: 301-312
... VLDB Journal 7(2): 96-114(1998); Markus Tresch, Neal Palmer, Allen
Luniewski: **Type Classification of Semi-Structured Documents**. VLDB ...
dbweb.csie.ncu.edu.tw/DBLP/dblp/ db/conf/sigmod/ConsensM94.html - 18k - Supplemental Result - Cached - Similar pages
[More results from dbweb.csie.ncu.edu.tw]

U. of Western Ontario /All Locations
... KL Tan, S. Dao, 251. **Type Classification of Semi Structured Documents**
/ M. Tresch, N. Palmer, A. Laniewski, 263. L/MRP: A Buffer Management ...
alpha.lib.uwo.ca:5701/search/agray+paul+a/agray+paul+a/ 19,-1,0,B/frameset&FF=agray+peter+m+d+1940&6,, - 22k - Supplemental Result - Cached - Similar pages

In order to show you the most relevant results, we have omitted some entries very similar to the 18 already displayed.

If you like, you can repeat the search with the omitted results included.

Google

Result Page: **Previous** 1 **2**

+"type classification of semi-structured documents" Google Search Search within results

CiteSeer
Scientific Literature Digital Library

Submit Documents Statistics About Feedback Help

Search Documents | Search Citations

Documents indexed by CiteSeer

Citations made by indexed documents

Most Cited Documents

Earth's largest free full-text index of scientific literature

NEC

A Robust System Architecture for Mining Semi-structured Data (1998) (Make Corrections) (5 citations)

Lisa Singh, Bin Chen, Rebecca Haight, Peter Scheuermann, Kiyoko Aoki

Knowledge Discovery and Data Mining

Abstract: The value of extracting knowledge from semi-structured data is readily apparent with the explosion of the WWW and the advent of digital libraries. This paper proposes a versatile system architecture for text mining that maintains structured data components in a relational database and unstructured concepts in a concept library. After a detailed explanation of our system architecture, we briefly describe IRIS, our prototype rule generation system Introduction Although much attention has been... (Update)

Context of citations to this paper: More

...For preconstrained textual mining, this structure facilitates the rule discovery process. 3. **2 Conceptual Model In our previous work [5], we proposed a system architecture that attempts to provide an infrastructure robust enough to facilitate the discovery of rules from**...

...but most of them regard doncument collection as semi structured or their purpose is to classify semi structured data. L. Singh et. **al [7][8] have attempted to discover knowledge from semi structured data.** However, in their approach, semi structured data means unstructured 7...

Cited by: More
IRIS: Our prototype rule generation system - Lisa Singh Peter (Correct)
Data Mining Architectures - A Comparative Study - Thomas, Jayakumar, Muthukumaran (Correct)
Web Mining Research: A Survey - Kosala, Blockeel (2000) (Correct)

Active bibliography (related documents): More All
0.4: An Algorithm for Constrained Association Rule Mining in.. - Lisa Singh (Correct)
0.3: Mining Is-Part-Of Association Patterns From Semistructured Data - Wang, Liu (Correct)
0.2: Generating Association Rules from Semi-Structured Documents.. - Lisa Singh (1997) (Correct)

Similar documents based on text: More All
0.2: Ozone: Integrating Structured and Semistructured Data - Lahiri, Abiteboul, Widom (2000) (Correct)
0.2: Storing Semistructured Data with STORED - Deutsch, Fernandez, Suciu (1999) (Correct)
0.2: Using Nested Tables for Representing and Querying.. - Silva, Filha.. (Correct)

Related documents from co-citation: More All
3: Fast Algorithms for Mining Association Rules - Agrawal, Srikant - 1994
3: Generating Association Rules from SemiStructured Documents Using an Extended Con.. - Singh, Scheuermann et al. - 1997
2: Querying semi-structured data - Abiteboul - 1997

BibTeX entry: (Update)

L. Singh, B. Chen, R. Haight, P. Scheuermann, and K. Aoki, "A robust system architecture for mining semi-structured data," In Proceedings of International Conference on Knowledge Discovery and Data Mining, 1998. http://citeseer.nj.nec.com/singh98robust.html More

```
@inproceedings{ singh98robust,
   author = "Lisa Singh and Bin Chen and Rebecca Haight and Peter Scheuermann and Kiyoko Aoki",
   title = "A Robust System Architecture for Mining Semi-Structured Data",
   booktitle = "Knowledge Discovery and Data Mining",
   pages = "329-333",
```

year = "1998",
url = "citeseer.nj.nec.com/singh98robust.html" }

Citations (may not include all citations):
971 Introduction to Modern Information Retrieval (context) - Salton, McGill - 1983 **Book Details from Barnes & Noble**
21 Discovering Trends in Text Databases - Lent, Agrawal et al. - 1997
10 Generating Association Rules from Semi-structured Documents .. - Singh, Scheuermann et al. - 1997
5 Type Classification of Semi-structured Documents (context) - Tresch, Palmer et al. - 1995
3 Mining Associations in the Presence of Background Knowledge (context) - Feldman, Hirsh - 1996
1 and Kohonen (context) - Lagus, Honkela et al. - 1996
1 Automatic Preprocessing & Transformation of Semi-Structured .. (context) - Singh - 1997

Documents on the same site (http://www.ece.nwu.edu/EXTERNAL/dbwww/research/mining/projects.html):

Generating Association Rules from Semi-Structured Documents.. - Lisa Singh (1997) (Correct)
An Algorithm for Constrained Association Rule Mining in.. - Lisa Singh (Correct)

Welcome to IEEE *Xplore*®

- Home
- What Can I Access?
- Log-out

Tables of Contents

- Journals & Magazines
- Conference Proceedings
- Standards

Search

- By Author
- Basic
- Advanced

Member Services

- Join IEEE
- Establish IEEE Web Account
- Access the IEEE Member Digital Library

IEEE ANNOUNCES NEW RELEASE FOR IEEE *XPLORE* ENHANCEMENTS - LEARN MORE.

IEEE *Xplore* provides full-text access to IEEE transactions, journals, magazines and conference proceedings published since 1988 plus select content back to 1950, and all current IEEE Standards.

FREE TO ALL: Browse tables of contents and access Abstract records of IEEE transactions, journals, magazines, conference proceedings and standards.

IEEE MEMBERS: Browse or search to access any complete Abstract record as well as articles from IEEE Spectrum Magazine. Access your personal online subscriptions using your active IEEE Web Account. If you do not have one, go to "Establish IEEE Web Account" to set up an account.

CORPORATE, GOVERNMENT AND UNIVERSITY SUBSCRIBERS: Search and access complete Abstract records and full-text documents of the IEEE online publications to which your institution subscribes.

Cookies Enabled
Click for more information

IEEE *Xplore* Quick Links

- New This Week
- OPAC Linking Information
- Email Alerts
- Your Feedback
- Technical Support
- No Robots Please
- Release Notes
- IEEE Online Publications

Current Activities

Clustering Binary Objects

Brian T. Luke

Binary objects are simply objects that are described by a finite-length bit string. They are often used in chemical databases because they can be very useful in searching for a particular compound. If the search string has a particular bit turned ON and a database compound does not, then that structure can be passed over. Two possible representations are Molecular Fingerprints and Structural Keys.

The bit strings that characterize two objects can also be used to calculate a "distance." This effective distance can then be used with a clustering algorithm to place the objects into groups. One of the most popular methods for calculating the distance between two objects is to first determine their similarity. If a general similarity metric is denoted S, it usually has a value between 0.0 (the bit strings have no ON bits in common) and 1.0 (the two bit strings are identical). A distance can then be given by $A(1-S)^M$, where M is a non-negative exponent and A is a constant. Conversely, each bit string can be thought of as components of an L-dimensional vector and the similarity as just the cosine of the angle between them (which is actually the case for the Ochini similarity). Therefore, a distance can be (A)arcos(S), or just a constant times the value of this angle.

If the bit string has a length of L, it is possible to go down this string and count the number of times a bit is ON in both strings, ON in one and OFF in the other, or OFF in both strings. The four sums are presented in the table below.

		Object *j*	
		0	**1**
Object *i*	**0**	B_{00}	B_{01}
	1	B_{10}	B_{11}

The first subscript refers to the value of the bit for object *i* and the second for object *j*, summed over all L bits. Therefore, B_{10} is the number of times a bit is ON in *i* and OFF in *j*.

Some other symbols that will be used in this section are

B_i ($= B_{10} + B_{11}$) is the total number of ON bits in object *i*.
B_j ($= B_{01} + B_{11}$) is the total number of ON bits in object *j*.
B_C ($= B_{11}$) is the total number of times a bit is ON in both bit strings.
B_I ($= B_{00} + B_{11}$) is the total number of times the two bit strings agree.
L ($= B_{00} + B_{01} + B_{10} + B_{11}$) is the length of the bit string.

With these definintions, commonly used similarity metrics are presented in the following table.

Label	Equation
Simple Matching Sokal & Michener	$SM = B_I/L$
Russel & Rao	$RR = B_C/L$
Tanamoto Coefficient	$TC = B_C/(B_i+B_j-B_C)$
Dice, Czekanowski, or Sorensen or Nei & Lie's genetic distance	$DCS = 2B_C/(B_i+B_j)$
Rogers & Tanamoto	$RT = B_I/(2L-B_I)$
Jaccard's Coefficient of Community	$JCC = B_C/(L-B_{00})$
Kulczynski	$KS = (1/2)(B_C/B_i+B_C/B_j)$
Ochini (cosine similarity function [1])	$CS = \frac{B_C}{\sqrt{B_i B_j}}$
R Package, v4.0	$S03 = 2B_I/(L+B_I)$ $S09 = 3B_C/(3B_C+B_{10}+B_{01})$ $S10 = B_C/(B_C+2B_{10}+2B_{01})$
Jaccard's Similarity	$JS_{i-j} = B_C/B_i$ $JS_{j-i} = B_C/B_j$
Tversky Index	$TI_{ab} = B_C/(aB_{10}+bB_{01}+B_C)$

The last two metrics are different from the rest. Jaccard's Similarity is different in that it is not symmetrical; the similarity of *i* to *j* is different than the similarity of *j* to *i* unless the two bit strings have exactly the same number of ON bits ($B_i=B_j$). This metric is useful if you want to measure a "substructure similarity." If *i* is a substructure of *j*, then every bit that is ON in *i* will also be ON in *j* though *j* will also have other bits ON. The other metrics will yield a similarity less than 1.0, while JS_{i-j} produces a similarity of 1.0. Similarly, if *j* has less bits ON than *i*, JS_{j-i} can be used to measure a substructure distance.

The Tversky Index is unique in that it has two adjustable parameters, *a* and *b*. If a=1 and b=0, $TI_{10}=JS_{i-j}$. If a+b=1, this expression can be rewritten as

$$TI_{ab} = B_C/(aB_i+bB_j)$$

and if they are both equal to one-half, this just becomes one-half of Sorensen's Index, SI. This also means that for particular values of *a* and *b*,

the Tversky Index may have to be scaled so that identical bit strings produce a similarity of 1.0.

A dissimilarity metric, D, is a measure of the difference between two bit strings. This metric has a value of 0.0 for identical bit strings and increases as they become less similar. It can therefore be scaled by any constant to yield a distance. A dissimilarity metric can easily be formed by taking D=(1-S), where S is a similarity value from the table above. Certain dissimilarities presented in the literature are given in the table below. They are proportional to either the sum or the product if the disagreement between the bit strings, B_{10} and B_{01} (note that $B_{10}+B_{01}=L-B_I$ is the Hamming Distance between these bit strings).

Label	Equation
Euclidean Dissimilarity	$DE = \sqrt{(L - B_I)/L}$
Squared Euclidean Dissimilarity	$DSE = (L-B_I)/L$
Pattern Difference	$DP = 4B_{10}B_{01}/L$
Lance & Williams Bray-Curtis nonmetric coefficient	$DLW = (L-B_I)/(B_i+B_j)$
Squared Euclidean Substructure Dissimilarity	$DSES_{i-j} = (B_i-B_C)/B_i = B_{10}/B_i$ $DSES_{j-i} = (B_j-B_C)/B_j = B_{01}/B_j$

In chemical databases, this metrics can yield a significant dissimilarity (distance) if a substructure of a molecule is compared to the full molecule. To remove this problem, Daylight Systems developed a Squared Euclidean Substructure Distance, DSES. If compound *i* is smaller than compound *j* (less bits turned ON), $DSES_{i-j}$ may yield a more accurate distance. Similariy, if object *j* has less bits turned ON than *i* ($B_j < B_i$), $DSES_{j-i}$ can be used.

With any of these distance metrics, a clustering algorithm can be used to place these binary objects into groups. Since it is not possible to calculate an average position between two binary objects (unless they are identical), one cannot determine a centroid for a group of binary objects. To circumvent this, a medoid [2] can be used instead. A medoid is simply a bit string that minimizes the sum of the distance to all objects in the group, and can be used with Wards Method (described in Agglomerative Linkages) and K-means clustering.

In 1987, Peter Willett [3] studied the clustering of chemical databases and found that Jarvis-Patrick clustering [4] produced the best results.

I used a similar procedure to try to cluster nucleotide sequences. Because of the large variation in the number of 1's for different bit strings, I found that the following distance metric produced better results.

$$D'_{SE} = B_j(B_i - B_c)/B_i^2$$

One final point is that it is possible to weight each bit in the string. If b_{ik} is the bit in position *k* (k=1,2,...L) for object *i*, it can be weighted by w_k. The terms in the above expressions become

$$B_i = \Sigma_{k=1}^{L} w_k b_{ik}$$

$$B_c = \Sigma_{k=1}^{L} w_k b_{ik} b_{jk}$$

$$B_j = \Sigma_{k=1}^{L} w_k \delta_{ijk}$$

$$\delta_{ijk} = 1 \text{ if } b_{ik} = b_{jk}$$

$$0 \text{ if } b_{ik} \neq b_{jk}$$

The weight for each bit can be related to the "inverse frequency" of being turned on. In other words, if this bit is on for all N objects to be clustered, it cannot be used as a descriminator. Conversely, if it is on in only a few of the objects, it may be an important descriminator and should be given a relatively large weight. One such inverse frequency weight is given by the expression

$$w_k = \log(N/f_k)$$

where f_k is the number of objects that have this bit turned on.

Note that the discussion above is limited to a comparison of bit strings that all have the same length. Clustering Character Objects describes metrics for comparing character strings, and character or binary strings of different length.

References:

[1] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, 1989.

[2] V. Guralnik and G. Karypis, "A Scalable Algorithm for Clustering Protein Sequences," in *Workshop on Data Mining in Bioinformatics*, (2001) 73-80.

[3] "Similarity and Clustering in Chemical Information Systems," P. Willett, Research Studies Press, Wiley, New York, 1987.

[4] R.A. Jarvis, E.A. Patrick "Clustering Using a Similarity Measure Based

on Shared Near Neighbors", IEEE Transactions on Computers, C22, 1025-1034 (1973).